99.95% accuracy - An accuracy measure usually used for key-entry or OCR, this number literally translates to the percentage of characters that are correct. 99.95% means that there are no more than 5 character errors per 10,000 characters, which for typical materials translates to 1-2 erroneous characters per page. 99.99% accuracy is 5 times as accurate with 1 error per 10,000 characters or 1 error every 5-10 pages. In DCL's electronic conversions, the standard character accuracy level is 100%.

Aggregator - A company who specializes in selling content from multiple sources via the Web. Generally, the aggregator's site is focused on a particular subject matter. Although aggregators are most common in the Scientific, Technical and Medical(STM) world, many are now popping up in other fields such as Libraries, Technology and Education.

Ambiguous Mapping - Ambiguous mapping occurs when a particular style, code or string maps to two or more possible SGML tags, depending on context or content. For example, italicized text may map to an SGML tag used to mark case names ("Smith v Jones"), an SGML tag used to mark foreign words ("c'est la vie"), or an SGML tag used solely for emphasis ("almost"). The number of such ambiguities can usually be resolved programmatically (e.g. italicized text with the word v is a case name).

ASP - An Active Server Page is an HTML page that includes scripts that are processed on the server side before the page is sent to the user. The primary purpose of using ASPs is so that a page can be tailored specifically to the user, based on his or her preferences. Basically the page pulls information from a database and then builds the final page on the fly before sending it to the browser. Examples of ASPs are "My Yahoo" and the customized pages that investment houses provide to allow you to view "your portfolio" as soon as you sign on.

CALS Tables - This model for the representation of tabular data was originally defined by the US Department of Defense as part of its CALS document interchange initiative. The table model (defined in military standard MIL-M-28001B) has become a de facto standard within the SGML industry.

Cascading Style Sheet - CSSs allow authors and users to attach style (e.g., fonts, spacing, and aural cues) to structured documents (e.g., HTML documents and XML applications). CSSs separate the presentation style of documents from the content of documents, and thereby simplify Web authoring and site maintenance. Both Netscape and IE now support CSSs.

CGM Computer Graphics Metafile is a graphics file format developed by experts working under the auspices of ISO and ANSI, and was designed specifically as a common format for the platform-independent interchange of raster (bitmap) and vector data. This format is used primarily to store vector

graphics information. CGM files typically contain either vector or raster data, but rarely both. Used in its primary role as a vector format, it offers the advantage of small file size and resolution independence, while not being tied to a specific software package or hardware platform. CGM was adapted by the Department of Defense as one of the CALS initiative standards.

Conditional Text - Conditional text allows the selective inclusion of a piece of text in an output document based on a series of conditions. A desktop publishing program which supports conditional text allows a user to have a one master document with a series of variant output documents. For example, a software manufacturer may want to distribute one user manual to its customers and deliver the same manual with additional text to its Technical Support people. Conditional text makes this possible. Packages that support conditional text include FrameMaker and Bookmaster.

DPI - Dots per inch is a measure of the sharpness or resolution in an image. Higher DPIs result in greater quality images although they can dramatically increase file size. The effect of this is that images will print more slowly or display more slowly on a computer screen. With the Internet, sophisticated compression algorithms have become popular to dramatically reduce file size without compromising quality. The JPEG format is an example of such compression. For web display 72DPI is typical, while for printing to a common laser printer 300 or 600 are more common. In Desktop Publishing, DPIs are typically much higher.

DTD - A document type definition is a specific definition that follows the rules of the Standard Generalized Markup Language (SGML). A DTD is a specification that accompanies a document and identifies markup codes, and the rules for their use. SGML documents need to be parsed or validated to ensure that they conform to the DTD. A DTD is optional with XML, but highly recommended with more complex document sets.

GIF - Graphics Interchange Format is the most common format for graphic images on the Internet. This highly-compressed format is used to display 2-dimensional raster images. A newer version, GIF 89a allows for an animated GIF, which is a short sequence of images within a single GIF file. GIF files are generally not used for photographs on the Web; JPEGs are optimized for that purpose.

The LZW compression algorithm used in the GIF format is owned by Unisys, and companies that make products that use the algorithm need to license its use from Unisys.

"Glass Typewriter" - This particular problem is very often an issue with data authored in the days preceding the sophisticated desktop publishing packages and word processors we know today. On older, proprietary document systems, data was often formatted inconsistently with the singular

goal that it appear correctly on screen. This “glass typewriter” approach is not uncommon, and while it served its function for display purposes, it greatly reduced the underlying structural integrity of the data. Most markedly, the practice greatly increases the complexity and effort of enhancing and converting data to more structured formats like XML, SGML, and FrameMaker.

HTML - Hypertext Markup Language is the set of "markup" codes or tags inserted in files intended for display on the World Wide Web. This markup tells the Web browser how to display a Web page's text and images. Examples of typical HTML tagging include the following:

<html><title>American Ski Association Welcome</title> <h1>The Joy Of Skiing</h1> <h4>by Jim Smith</h4> <h2>Introduction</h2> Skiing is one of the fastest growing sports in America. This book is a tribute to the sport and a how-to guide to getting started. We hope that you enjoy it, and get out on the slopes real soon! Note: All opinions expressed in this book belong to the author. </html>

HTML is a standard recommended by the World Wide Web Consortium (W3C) and adhered to by the major browsers.

IETM - Interactive Electronic Technical Manual. This technical manual is usually stored on CD-ROM and provides for unique user interactivity. In general, the IETM helps do away with the page-turning that is normally associated with paper manuals in order to see referenced figures, tables, chapters, etc and to do trouble-shooting. In the case of referenced figures and tables, etc., the IETM lets the user hyperlink directly to the referenced item. In a trouble-shooting section, the user simply clicks on the current problem and the IETM walks him/her through the trouble-shooting process by specifying a trouble-shooting test and the possible results of the test.

JPEG - Joint Photographic Experts Group files are used for monochrome, gray scale or full-color digital still images. JPEGs use compression to tremendously decrease file size while still maintaining high image quality. JPEG has become the de facto standard for photographs on the Web.

Mapping - In the context of XML/SGML conversions, this means the specification of the SGML tagging to be produced when a particular style (paragraph or font), coding, or string of text is found in the input file. For example, the ChapTitle style may map to the SGML tagging <chapter><title>...</title>, meaning that when the paragraph style ChapTitle is found in the input file, then the SGML-encoding software will produce <chapter><title>...</title> with the "..." representing the text found in the paragraph styled as "ChapTitle".

Master Format - In DCL's conversion methodology, this is a format into

which all incoming data is converted in order to standardize it for further conversion processing. DCL's master format uses SGML as its base. From here, data can be converted to multiple output formats, and even to multiple DTDs. The major advantage of this approach is that all incoming formats can be normalized into a common dataset on which DCL's conversion software can operate. The approach also facilitates multi-purposing of the same data for multiple output formats.

MathML - The Mathematical Markup Language, is an XML based language used for displaying mathematical notation and content, especially on the web. It is a World Wide Web Consortium (W3C) recommended standard, and has been receiving increasing support by mathematical software vendors.

OCR - Optical Character Recognition is a visual recognition process that turns printed or written text into an electronic character based file. The process involves photo-scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, typically ASCII. In OCR processing, the page image is scanned, then analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. Popular commercial OCR packages include the Xerox company's *TextBridge* and Adobe's *Acrobat Capture*.

Parse - While traditionally a concept of syntax and grammar validation, when used in relation to mark-up languages, this terms refers to a process of validating files by checking that tags are applied legally according to a pre-defined structure. This structure is typically defined by the Document Type Definition (DTD). Common terms used in mark-up validation are "parser" (a piece of software that validates) and "parsed".

PDF - Portable Document Format ("PDF") reproduces the documents almost precisely as they were originally composed, provides built-in compression, is supported by all popular operating systems and is compatible with most printers. The freely available Adobe Acrobat Reader is required to view, print and search PDF documents. The PDF format was developed by Adobe, is modeled after the PostScript language, and is both device and resolution independent.

While mark-up languages are generally preferred for content-oriented materials, PDF files are especially useful for documents where appearance is critical. A PDF file contains one or more page images, each of which you can zoom in on or out from.

Raster - Also referred to as bitmap images, these are images that are represented by a sequence of pixels (picture elements) or points, which when taken together, describe the display of an image on an output device. There are many different raster image formats in use, among them GIF, JPEG, PCX, and TIFF.

Resolution - Resolution refers to the number of pixels (individual points of color) contained on a display monitor. The number is expressed in terms of the number of pixels on the horizontal axis and the number on the vertical axis. The sharpness of the image on a screen depends on both the resolution and the size of the monitor. The same pixel resolution will gradually lose sharpness as monitor size increases because the same number of pixels are now being spread over a larger physical area. Resolution is similar to DPI except that DPI is more typically used in regards to printed output.

Sample Markup - An initial step in the Proof of Concept phase, this refers to the text of a sample document with the SGML tags inserted. The sample markup may be a hardcopy document with the tags written in or it may be an electronic SGML file along with the corresponding hardcopy.

SGML - Standard Generalized Markup Language is an internationally agreed standard for information representation. SGML can be used for publishing in its broadest definition - from single medium conventional publishing on paper to on-line multi-media database publishing. SGML can be used to produce files which can be read by people, and exchanged between machines and applications in a straightforward manner.

Styled - Most modern word processing and desktop publishing programs allow the user to supply a base stylesheet (sometimes called a template) so that 'like' paragraphs can all have a similar look. A document is called 'styled' if the component paragraphs are produced by use of these styles.

Stylesheet - A master document template made up of a collection of styles. Most desktop publishing and word processing packages come with a standard stylesheet (also called template) that includes styles for things such as first-level headings and bulleted list items. Stylesheets are critical to enforcing structure and consistency across document sets, especially where multiple authors are involved.

Template - see stylesheet.

Text Frames - Text Frames are popular in desktop publishing, and are used to position text absolutely on a page. Many of the popular magazines that you read render sidebars and the like by using text frames. Text frames or boxes can significantly complicate the conversion process because they do not follow the logical 'story' structure of the document.

TIFF - Tag Image File Format is a common format for exchanging raster (bitmapped) images between application programs. Usually identified with the ".tiff" or ".tif" filename extension, the format was developed in 1986 by an industry committee chaired by the Aldus Corporation (now part of Adobe). Microsoft and Hewlett-Packard were also on the committee. One of the more common image formats, TIFFs are common in desktop publishing, faxing,

and medical imaging applications.

Unstyled - Unstyled documents are produced by using specific text formatting (such as justification, emphasis, tabs, indents, and font selection) for each paragraph individually, rather then by giving them a specific appearance based on selection of a particular style from a preselected stylesheet. This approach undermines the structural integrity of a document and often leads to inconsistency within a set of documents. Unstyled materials add tremendously to the task of performing large-scale automated conversions.

Vector - Vector images are images that are represented by collections of independent line and shape objects which are typically defined by mathematical formulas. This makes these images easier to modify than raster images. Popular vector image programs include Adobe Illustrator, CorelDraw, and AutoCad. Typically, each program will have its own vector file format.

WYSIWYG (pronounced "wiz-ee-wig") - Literally, What-You-See-Is-What-You-Get, this refers to an editor or program that incorporates a graphical user interface (GUI) so that a developer (usually working with code or markup) can see the end result while creating the document. Many products now exist for web design that allow pages to be build graphically without the user having an in-depth knowledge of the underlying HTML code. Adobe's PageMill and Microsoft's Front Page are such products.

XML - Extensible Markup Language is a subset of ISO 8879, Standard Generalized Markup Language (SGML). XML has been designed specifically to function on the Web, and both major browsers support it. Currently a formal recommendation from the World Wide Web Consortium (W3C), XML is similar to HTML in that both XML and HTML contain markup symbols to describe the contents of a page or file. HTML, however, describes the content of a Web page only in terms of how it is to be displayed. XML describes the content in terms of what the data is that is being described. For example the <authname><affil> tags could indicate that the data following it was an author's name and his affiliation. This allows an XML file to be processed purely as data by a program as well as being displayed in a certain way. XML is "extensible" because, unlike HTML, the markup symbols are unlimited and self-defining.

XSL - Extensible Stylesheet Language is a stylesheet language that gives us the ability to specify how data coded with XML will format on screen. This language was developed based on the ISO companion standard for SGML known as DSSSL (Document Style Semantics and Specification Language.)